

2002

Sample design and estimation for household surveys

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/theses>

University of Wollongong

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Recommended Citation

Clark, Robert Graham, Sample design and estimation for household surveys, Doctor of Philosophy thesis, School of Mathematics and Applied Statistics, University of Wollongong, 2002. <https://ro.uow.edu.au/theses/2055>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

SAMPLE DESIGN AND ESTIMATION FOR HOUSEHOLD SURVEYS

A thesis submitted in fulfilment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

from

UNIVERSITY OF WOLLONGONG

by

Robert Graham Clark

SCHOOL OF MATHEMATICS AND APPLIED STATISTICS

2002

Abstract

Household surveys are a widely-used tool for obtaining information about a population of people. A sample of households is selected followed by a sample of people within selected households. Households exhibit structure with variables measured on people in the same household often being dependent. Household sizes vary significantly and the strength of dependencies within a household may depend on its size. Traditional sample design and estimation methods ignore these dependencies.

Methodologies which explicitly allow for the dependencies which may arise within households are developed in four areas:

- estimating the design effects of standard estimators, for survey design;
- constructing new estimators to exploit dependencies within households;
- selecting a set of auxiliary variables to use in regression estimation;
- allocating the sample sizes of households and of people within households.

In each case, new methods will be developed which allow for the population structure of people within households. The new methods will be compared theo-

retically and numerically to existing methods to show whether, and under what conditions, it is worthwhile explicitly allowing for this population structure.

This thesis is concerned with the sampling error of estimators of population totals; that is, the error due to selecting only a sample and not the whole population of people. The model-assisted framework will be used.

The thesis finds that the population structure of people within households can be exploited to give several useful innovations. It is shown that, in estimating the variance at the sample design stage, the variation of household size must be considered. This variation is ignored in existing methods for estimating the design effect, and a more accurate method is developed. It is found that minor improvements can be made to standard estimators of total by considering within-household dependencies. An “integrated weighting” method, based on a linear contextual model, which has important practical advantages is found to often have slightly lower variance than non-integrated methods, contrary to common belief. Existing criteria for selecting which auxiliary variables to use in regression estimation are extended to the case of two-stage sampling, and applied to household surveys. In most household surveys, either one person or all people are selected from each selected household. More general designs, in which the number of people selected is a function of the number of people in the household, are developed. The fact that the number of people in a household is small leads to some novel and efficient sample designs and estimators.

Certification

I, Robert Graham Clark, declare that this thesis is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Robert Clark

4 March 2002

Acknowledgements

I would like to thank my friends and colleagues who helped to make my Ph.D. a fulfilling and enjoyable experience.

I am grateful to my supervisor, Professor David Steel, for his encouragement, ideas and good advice. David also gave me excellent opportunities to travel and collaborate with other researchers in survey statistics.

This research was funded by the Australian Research Council and the Australian Bureau of Statistics(ABS). Mr Frank Yu from the ABS supported my leave of absence from the ABS and provided valuable advice during the Ph.D. Both Frank Yu and Professor Alan Welsh from the University of Southampton helped to develop the research plan.

The Department of Social Statistics at the University of Southampton kindly hosted me for one year and provided an active and friendly research environment. Professor Ray Chambers made me welcome at Southampton and generously gave his time, ideas and hospitality.

Thanks to my friends at the University of Wollongong and the University of Southampton for sharing many enjoyable and challenging experiences.

Thank you Carolyn Silveri, Kerrie Gamble, Anne Owens, Kathy Hooper, Debbie Edwards and the University of Wollongong Office of Research for providing administrative support.

Thanks Bill, Stephanie, Alistair, Matt and Kathy for giving me homes away from home in Canberra. Thanks Patricia and Michelle for looking after my house.

Bill Allen, Eric Beh, Russell Familiar, Michael Meagher, Nick von Sanden and Craig McLaren were a great help in proofreading the thesis.

I am grateful to my parents for their support and example.

Contents

1	Introduction	1
1.1	The Definition and Relevance of Household Surveys	1
1.2	Theoretical Framework	4
1.3	Scope of the Thesis	6
1.4	Special Features of Household Surveys	8
1.5	Problems That Will Be Considered	10
1.6	Structure of the Thesis	16
2	Literature Review	19
2.1	Notation	19
2.2	The Model-Assisted Framework	22
2.2.1	Design-Based and Model-Based Expectations	22
2.2.2	Asymptotics	26
2.2.3	Regression Estimators	27
2.2.4	Lower Bounds for the Anticipated Variance	32
2.2.5	Other Approaches to Finite Population Inference	37
2.3	Design Effects for Two-Stage Surveys	40

2.3.1	The Relevance of Design Effects	40
2.3.2	Variance of Inverse Probability and Hájek Estimators . . .	42
2.3.3	Decompositions of the Design Effect	44
2.4	Selecting Auxiliary Variables to use in Estimation	51
2.4.1	Introduction	51
2.4.2	Model-Assisted Methods for SRSWOR	55
2.4.3	Prediction Approach	58
2.4.4	Relevance to Household Surveys	62
2.5	Allocating Sample to Strata and Stages	63
2.5.1	The Neymann Allocation Problem	63
2.5.2	Allocation to Strata and Stages	65
2.5.3	Sample Design for Household Surveys	69
3	Design Effects for Household Sampling	73
3.1	Introduction	73
3.2	Design Effects of Estimators of Total for Household Sampling . .	76
3.2.1	Comparison of Hájek and Inverse Probability Estimators .	76
3.2.2	Decomposing the Design Effects of Simple Estimators of Total	79
3.2.3	Definition of Intra-PSU Correlation	84
3.2.4	Design Effects for Regression Estimators and Estimators of Regression Parameters	87

3.3	Model-Assisted Approaches	92
3.3.1	Approximating the Design Effect using Orthogonality Assumptions	92
3.3.2	Approximating the Design Effect using a Model	93
3.3.3	Using Models to Compare the Design Effects of Hájek Estimator, Regression Estimator and Estimator of Regression Coefficient	95
3.4	Using Design Effects for Survey Design	99
3.5	Including Children in Scope	107
3.6	Summary of Chapter 3	109
4	Estimators Exploiting Within-Household Dependencies	111
4.1	Introduction	111
4.2	Model-Assisted Optimality	116
4.2.1	Introduction	116
4.2.2	Lower Bounds for the Anticipated Variance	117
4.2.3	Correlations between Y_i and Y_j	126
4.3	Estimators based on the Linear Contextual Model	128
4.3.1	The Linear Contextual Model	128
4.3.2	Contextual GREGs	130
4.3.3	The All/Household Design and Integrated Weighting	133
4.3.4	Comparing the Efficiency of $\hat{T}_{C\pi}$ and $\hat{T}_{r\pi}$	136

4.4	Simulation Study of Person and Contextual GREGs	143
4.5	Summary of Chapter 4	148
5	Selecting Auxiliary Variables for Household Surveys	151
5.1	Introduction	151
5.2	Methods Assuming Equal Variances and Zero Covariances	156
5.3	Methods Allowing Unequal Variances and Nonzero Covariances .	164
5.4	Asymptotic Unbiasedness of the Ultimate Cluster Variance Criterion	172
5.4.1	Model-Based Asymptotics	172
5.4.2	Magnitude of the Difference of the Model Mean Squared Errors	174
5.4.3	Asymptotic Unbiasedness	176
5.5	Combining Model-Based and Design-Based Methods	177
5.5.1	Motivation	177
5.5.2	Combined Method	179
5.6	Design Expectation of the Model Mean Squared Error for a Simple Case	182
5.7	Simulation Study of Selection of Auxiliary Variables	188
5.8	Summary of Chapter 5	195
6	Allocating Sample Sizes of People and Households	199
6.1	Introduction	199
6.2	Cost Models and Variance	203

6.3	Simple Rounded Allocations	211
6.3.1	Unstratified First Stage Design	211
6.3.2	Stratified Design	213
6.3.3	Proportional Design	215
6.4	Best Integer Allocations	216
6.4.1	Unstratified Design	216
6.4.2	Stratified Design	217
6.4.3	Proportional Design	218
6.5	Fractional Allocations	219
6.5.1	Allowing Within-PSU Sample Sizes to be Random	219
6.5.2	Estimators of Total for Fractional Allocations	221
6.5.3	Cost and Variance for Fractional Allocations	227
6.5.4	Outline of Derivation of Optimal Fractional Allocation	229
6.6	Weighting for Fractional Designs	230
6.6.1	Optimal Weights	230
6.6.2	Special Cases	235
6.7	Optimal Fractional Allocations	240
6.8	Numerical Study of Within-Household Sampling Schemes	246
6.9	Summary of Chapter 6	259
7	Conclusion	263
7.1	Summary and Conclusions	263

7.2 Further Research	268
A Proofs and Additional Tables for Chapter 2	273
A.1 Derivation of (2.17), (2.18), (2.19), (2.20) on Page 43	273
A.2 Derivation of (2.34) on Page 56	274
B Proofs and Additional Tables for Chapter 3	277
B.1 Proof of Theorem 3.1: Decompositions of Deffs for Household Sam- pling	277
B.2 Proof of Theorem 3.2: Orthogonality Approximations for Deffs . .	278
B.3 Proof of Theorem 3.3: Model Expectations of Deffs	280
B.4 Proof of Theorem 3.4: Some Model Expectations	282
B.5 Proof of Lemma 3.5: Decompositions for PSUs of the Same Size .	284
B.6 Proof of Theorem 3.6: Decompositions of C_{NY2} and C_{NS2}	285
B.7 Proof of Lemma 3.7: A Property of the Intraclass Correlation . .	285
B.8 Tables with Children Included	287
C Proofs and Additional Tables for Chapter 4	293
C.1 Proof of Theorem 4.1: Godambe-Joshi Lower Bound with Random X	293
C.2 Proof of Corollary 4.2: Optimal Estimation for One/Household Sampling	295
C.3 Proof of Corollary 4.3: Optimal Estimation for Constant Variances and Covariances within PSUs	296

C.4	Proof of Theorem 4.5: Decomposition of the GREG	299
C.5	Proof of Theorem 4.6: Contextual GREGs for All/Household Sam- pling	301
C.6	Proof of Theorem 4.7: AVs of GREGs under the Contextual Model	306
D	Proofs and Additional Tables for Chapter 5	309
D.1	Proof of Corollary 5.2: Expressing Δ in terms of the CVs of the Weights	309
D.2	Proof of Lemma 5.4: $\sum_{i \in U} \mathbf{k}^T \hat{\mathbf{u}}_i$ is a Regression Estimator	311
D.3	Proof of Theorem 5.5: Asymptotic Order of Δ	312
D.4	Proof of Theorem 5.6: Asymptotic Bias of $\hat{\Delta}_{ucv}$	314
D.5	Proof of Theorem 5.7: $\hat{\Delta}_{ucv}$ is first order Design-Unbiased	317
D.6	Proof of Corollary 5.8: Model MSE where Variances and Covari- ances are Equal	321
D.7	Proof of Theorem 5.9: Design Expectation of Model Variance of $\hat{T}_{r\pi}$	323
D.8	Proof of Corollary 5.10: Design Expectation of Model MSEs	328
D.9	Proof of Corollary 5.11: Design Expectation of Model MSEs for Linear Contextual Model	330
E	Proofs and Additional Tables for Chapter 6	335
E.1	Derivation of (6.9)	335
E.2	Proof of Lemma 6.1: First Order Taylor Series for \hat{T}_w	336
E.3	Proof of Theorem 6.2: Design Variance where n_g are Random . . .	338

E.4	Proof of Theorem 6.3: Weighting where n_g are Random	345
E.5	Proof of Result 6.5: Optimal Integer Distribution of n_g for Given θ_a	346
E.6	Proof of Theorem 6.6: The Optimal θ_a are Integers for the Strati- fied First Stage Design	350
E.7	Extra Tables: Sample Designs for Various Synthetic Variables and Cost Functions	352
F	Reports Produced from this Thesis	365
	References	367

List of Figures

- 6.1 Optimal Weights $w_{g(opt)}$ for Different R, θ_a 257
- 6.2 Second Stage Component of the Design Variance for Alternative
Weights w_g 258
- 6.3 Design Variance of $\hat{T}_{w(opt)}$ as a Function of θ_a 260

List of Tables

3.1	Decomposing Design Effect of \hat{T}_1 for All/HH Design	83
3.2	Decomposing Design Effect of \hat{T}_1 for One/HH Design	84
3.3	Decomposing Design Effect of $\hat{T}_{r\pi}$ for All/HH Design	90
3.4	Decomposing Design Effect of $\hat{T}_{r\pi}$ for One/HH Design	90
3.5	Decomposing Design Effect of $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ (sex contrast) for All/HH De- sign	91
3.6	Decomposing Design Effect of $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ (sex contrast) for One/HH Design	91
3.7	Comparison of the Deffs of Different Regression Estimators	97
3.7	(continued)	98
3.8	Breakdown of Employment Variable by Household Size	102
3.9	Breakdown of Income Variable by Household Size	102
3.10	Breakdown of Unemployment Variable by Household Size	103
3.11	Breakdown of FT Student Variable by Household Size	103
3.12	Breakdown of “Health Fair or Poor” Variable by Household Size .	104
3.13	Decomposition of C_{NY2}	105
3.14	Design Effect Approximations	106

3.15	Design Effect of \hat{T}_1 for Alternative Designs	107
4.1	Percent CVs of Contextual and Person GREGs by Design	145
4.2	$CV\% [\hat{T}_{C(BLU)}]$ for Different Choices of ρ	147
4.3	Decomposing the Variance Improvement from the Contextual GREG	147
4.4	Median Deffs of Subpopulation Estimates	148
5.1	Number of Auxiliary Variables Selected using UCV(employment)	191
5.2	Variance of Saturated GREGs and GREGs using UCV(employment) relative to $var_p [\hat{T}_1]$	193
5.3	Variance of Saturated GREGs and GREGs using Three Criteria (employment) relative to $var_p [\hat{T}_1]$	194
5.4	Variance of Saturated GREGs and GREGs using Three Criteria (English second language) relative to $var_p [\hat{T}_1]$	194
5.5	Variance of GREGs using UCV selection for Four Y Variables, relative to $var_p [\hat{T}_1]$ (m=250)	195
5.6	Variance of GREGs using UCV selection for Four Y Variables, relative to Saturated GREG (m=250)	195
6.1	Various Sample Designs, Employment Variable, $\bar{C}_{1a} = 1, \bar{C}_{2a} = 1$.	251
6.2	Various Sample Designs, Employment Variable, $\bar{C}_{1a} = 0.5, \bar{C}_{2a} = 1$	251
6.3	Various Sample Designs, Employment Variable, $\bar{C}_{1a} = 0.25, \bar{C}_{2a} = 1$	252

6.4	Optimal Integer Designs for Many Variables: Variance / Vari-	
	ance(all/hh)	253
6.5	Optimal Random n_g Designs for Many Variables: Variance / Vari-	
	ance(all/hh)	254
6.6	Variance(one/hh) / Variance(all/hh) for Various Synthetic Vari-	
	ables and Cost Models	254
6.7	Variance(best integer design) / min(Variance(one/hh), Variance(all/hh)	
	for Various Synthetic Variables and Cost Models	255
6.8	Variance (best random n_g) / Variance(best integer design) for Var-	
	ious Synthetic Variables and Cost Models	255
B.1	Decomposing Design Effect of \hat{T}_1 for All/HH Design	287
B.2	Decomposing Design Effect of \hat{T}_1 for One/HH Design	287
B.3	Decomposing Design Effect of $\hat{T}_{r\pi}$ for All/HH Design	288
B.4	Decomposing Design Effect of $\hat{T}_{r\pi}$ for One/HH Design	288
B.7	Comparison of the Deffs of Different Regression Estimators	288
B.7	(continued)	289
B.5	Decomposing Design Effect of $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ (sex contrast) for All/HH De-	
	sign	290
B.6	Decomposing Design Effect of $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ (sex contrast) for One/HH Design	290
B.8	Breakdown of Employment Variable by Household Size	291
B.9	Breakdown of Income Variable by Household Size	291

B.10	Breakdown of Unemployment Variable by Household Size	291
B.11	Breakdown of FT Student Variable by Household Size	292
B.12	Breakdown of “Health Fair or Poor” Variable by Household Size .	292
E.1	Synthetic Variable $R = 0.05, C_{1a} = 0.1, C_{2a} = 1$	352
E.2	Synthetic Variable $R = 0.05, C_{1a} = 0.25, C_{2a} = 1$	352
E.3	Synthetic Variable $R = 0.05, C_{1a} = 0.5, C_{2a} = 1$	353
E.4	Synthetic Variable $R = 0.05, C_{1a} = 0.75, C_{2a} = 1$	353
E.5	Synthetic Variable $R = 0.05, C_{1a} = 1, C_{2a} = 1$	354
E.6	Synthetic Variable $R = 0.1, C_{1a} = 0.1, C_{2a} = 1$	354
E.7	Synthetic Variable $R = 0.1, C_{1a} = 0.25, C_{2a} = 1$	355
E.8	Synthetic Variable $R = 0.1, C_{1a} = 0.5, C_{2a} = 1$	355
E.9	Synthetic Variable $R = 0.1, C_{1a} = 0.75, C_{2a} = 1$	356
E.10	Synthetic Variable $R = 0.1, C_{1a} = 1, C_{2a} = 1$	356
E.11	Synthetic Variable $R = 0.2, C_{1a} = 0.1, C_{2a} = 1$	357
E.12	Synthetic Variable $R = 0.2, C_{1a} = 0.25, C_{2a} = 1$	357
E.13	Synthetic Variable $R = 0.2, C_{1a} = 0.5, C_{2a} = 1$	357
E.14	Synthetic Variable $R = 0.2, C_{1a} = 0.75, C_{2a} = 1$	358
E.15	Synthetic Variable $R = 0.2, C_{1a} = 1, C_{2a} = 1$	358
E.16	Synthetic Variable $R = 0.3, C_{1a} = 0.1, C_{2a} = 1$	358
E.17	Synthetic Variable $R = 0.3, C_{1a} = 0.25, C_{2a} = 1$	359
E.18	Synthetic Variable $R = 0.3, C_{1a} = 0.5, C_{2a} = 1$	360

E.19 Synthetic Variable $R = 0.3, C_{1a} = 0.75, C_{2a} = 1$	360
E.20 Synthetic Variable $R = 0.3, C_{1a} = 1, C_{2a} = 1$	361
E.21 Synthetic Variable $R = 0.4, C_{1a} = 0.1, C_{2a} = 1$	361
E.22 Synthetic Variable $R = 0.4, C_{1a} = 0.25, C_{2a} = 1$	361
E.23 Synthetic Variable $R = 0.4, C_{1a} = 0.5, C_{2a} = 1$	362
E.24 Synthetic Variable $R = 0.4, C_{1a} = 0.75, C_{2a} = 1$	362
E.25 Synthetic Variable $R = 0.4, C_{1a} = 1, C_{2a} = 1$	362
E.26 Unemployment Variable, $\hat{T}_{r\pi}, C_{1a} = 0.5, C_{2a} = 1$	363
E.27 Various Sample Designs, Income Variable, $\hat{T}_{r\pi}, C_{1a} = 0.5, C_{2a} = 1$	363
E.28 Variance(best integer design) / Variance(all/hh) for Various Syn- thetic Variables and Cost Models	363
E.29 Variance(best integer design) / Variance(one/hh) for Various Syn- thetic Variables and Cost Models	364

